

Robust Bayesian Inference of Networks Using Dirichlet t -Distributions

Michael Finegold

mfinegol@andrew.cmu.edu

Department of Statistics, Carnegie Mellon University, Pittsburgh, U.S.A.

Mathias Drton

drton@uchicago.edu

Department of Statistics, The University of Chicago, Chicago, U.S.A.

July 6, 2012

Abstract

Bayesian graphical modeling provides an appealing way to obtain uncertainty estimates when inferring network structures, and much recent progress has been made for Gaussian models. These models have been used extensively in applications to gene expression data, even in cases where there appears to be significant deviations from the Gaussian model. For more robust inferences, it is natural to consider extensions to t -distribution models. We argue that the classical multivariate t -distribution, defined using a single latent Gamma random variable to rescale a Gaussian random vector, is of little use in highly multivariate settings, and propose other, more flexible t -distributions. Using an independent Gamma-divisor for each component of the random vector defines what we term the alternative t -distribution. The associated model allows one to extract information from highly multivariate data even when most experiments contain outliers for some of their measurements. However, the use of this alternative model comes at increased computational cost and imposes constraints on the achievable correlation structures, raising the need for a compromise between the classical and alternative models. To this end we propose the use of Dirichlet processes for adaptive clustering of the latent Gamma-scalars, each of which may then divide a group of latent Gaussian variables. Dirichlet processes are commonly used to cluster independent observations; here they are used instead to cluster the dependent components of a single observation. The resulting Dirichlet t -distribution interpolates naturally between the two extreme cases of the classical and alternative t -distributions and combines more appealing modeling of the multivariate dependence structure with favorable computational properties.

Key Words: Bayesian inference, Dirichlet process, graphical model, Markov chain Monte Carlo, t -distribution.

1 Introduction

Consider a random vector $\mathbf{Y} = (Y_1, \dots, Y_p)$ and an undirected graph $G = (V, E)$ with vertex set $V = \{1, \dots, p\}$. The Gaussian graphical model given by the graph G assumes that \mathbf{Y} follows a multivariate normal distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ may be any mean vector in \mathbb{R}^p but $\boldsymbol{\Sigma}$ is a positive definite covariance matrix that is constrained such that two variables Y_j and Y_k are conditionally independent given all the remaining variables $\mathbf{Y}_{\setminus\{j,k\}}$ whenever $\{j, k\}$ is not an edge in E . The conditional independence holds if and only if $\Sigma_{jk}^{-1} = 0$; see e.g. Lauritzen (1996). Hence, the graph of a Gaussian graphical model can be inferred from data by inferring the pattern of non-zero entries in the precision matrix $\boldsymbol{\Sigma}^{-1}$.

Different Bayesian methods have been developed for an uncertainty assessment in inference of the graph. Giudici and Green (1999), for instance, use a uniform prior on decomposable graphs and place a Hyper Inverse Wishart prior on the covariance matrix, which allows for exact local computation (Dawid and Lauritzen, 1993). In particular, a closed form marginal likelihood permits treatment of high-dimensional datasets (Dobra et al., 2004; Jones et al., 2005). Exact computations for non-decomposable graphs are much more involved (Roverato, 2002; Dellaportas et al., 2003; Atay-Kayis and Massam, 2005); for an approximate treatment see Lenkoski and Dobra (2011). Other recent literature providing different extensions to the basic Gaussian model includes Rajaratnam et al. (2008) and Carvalho and Scott (2009).

While appealing in many respects, Gaussian methods are susceptible to great impact of measurement errors. There is a substantial literature on robustness in Bayesian inference, including De Finetti (1961) and West (1984), but little work that directly targets graphical models. One commonly taken approach replaces multivariate normal by multivariate t -distributions (with univariate t -margins). T -distributions yield reasonable models for heavy-tailed marginal behavior and have the capacity to maintain good statistical efficiency when data are in fact Gaussian. Moreover, convenient closed form conditional distributions are available for use in Markov chain Monte Carlo algorithms that exploit the representation of t -random variables as ratios involving Gaussian and Gamma random variables.

The classical multivariate t -distribution can be defined in terms of an unobserved Gaussian random vector and a single unobserved Gamma divisor. In the context of this paper, modeling assumptions then refer to the latent Gaussian vector. Penalized likelihood techniques for graphical model selection with this classical t -distribution were explored in Yuan and Huang (2009). Highly multivariate data, however, often have pockets of contamination in many observations creating a scenario to which the classical t -distribution is poorly suited. This paper addresses this problem by developing methods based on more flexible t -distributions. The distribution we term the alternative t -distribution has independent Gamma scalars for each component of the latent Gaussian vector. This construction has been used in a frequentist treatment of graphical models (Finegold and Drton, 2011), but seems to have received little attention otherwise. While better suited to higher-dimensional analysis, the distribution’s use comes at increased computational cost and imposes constraints on the achievable correlation structures; see Fig. 9 in Finegold and Drton (2011). It is thus desirable to achieve a trade-off between the two extremes, ‘classical’ versus ‘alternative’.

The key new contribution of this paper is an adaptive method to interpolate between the classical and the alternative case. Our proposal uses Dirichlet processes to cluster the Gamma divisors associated with the components of the latent Gaussian vector. The common Gamma-value associated with a cluster is then used to divide all the associated Gaussian variables. The Dirichlet process clustering thus pertains to the possibly dependent components of a single multivariate observation, as opposed to the common case of clustering different independent observations. Compared to the alternative case, the Dirichlet t -proposal alleviates constraints on correlations and reduces computational effort when a small number of divisors is sufficient. While we are concerned with graphical models, there is no limitation to the use of the Dirichlet t -framework in other problems of multivariate statistics.

The paper is organized as follows. Section 2 lays out the Gaussian setup upon which our extensions are based. In Section 3, we describe Bayesian inference of network structures with classical and alternative t -distributions. The Dirichlet t -distribution is developed in Section 4. Numerical experiments (Section 5) and an analysis of gene expression data (Section 6) demonstrate that our new framework is computationally tractable and statistically efficient across a broad spectrum of data with outliers. A conclusion is given Section 7.

2 A Bayesian Gaussian Graphical Model

2.1 Background

Let $IW_p(m, \Phi)$ denote the Inverse Wishart (IW) distribution with degrees of freedom $m > p - 1$ and a positive definite scale matrix Φ . This distribution is supported on the cone of $p \times p$ positive definite matrices and has density

$$p(\Sigma \mid m, \Phi) = \frac{|\frac{\Phi}{2}|^{\frac{m}{2}}}{\Gamma_p(\frac{m}{2})} |\Sigma|^{-\frac{m+p+1}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(\Phi \Sigma^{-1}) \right\}, \quad (2.1)$$

where the determinant of a matrix \mathbf{A} is denoted $|\mathbf{A}|$ and

$$\Gamma_p(\alpha) = \pi^{p(p-1)/4} \prod_{i=1}^p \Gamma \left(\alpha - \left(\frac{i-1}{2} \right) \right)$$

is the multivariate gamma function with argument $\alpha > (p-1)/2$. The distribution is the conjugate prior for the covariance matrix of a multivariate normal distribution.

Let $G = (V, E)$ be a decomposable graph with vertex set $V = \{1, \dots, p\}$, and suppose $C_1, S_2, C_2, \dots, S_m, C_m$ is a perfect sequence of the graph's cliques C_i and separators S_i . Here and throughout the paper, we assume familiarity with basic graphical concepts as introduced in Lauritzen (1996). Let $M^+(G)$ be the set of positive definite matrices Σ with entry $\Sigma_{jk} = 0$ whenever $\{j, k\}$ is not an edge in E . For Gaussian graphical modeling, one needs a constrained version of the IW distribution for the covariance matrix Σ that has support such that the precision matrix Σ^{-1} lies in $M^+(G)$. The relevant distribution is known as the Hyper IW distribution and we denote it by $HIW(G, \delta, \Phi)$, where $\delta > 0$ is a degrees of freedom parameter

and Φ is a positive definite scale matrix. The distribution's density

$$f(\Sigma \mid G, \delta, \Phi) = \frac{\prod_{i=1}^m p(\Sigma_{C_i C_i} \mid \delta + |C_i| - 1, \Phi_{C_i C_i})}{\prod_{i=2}^m p(\Sigma_{S_i S_i} \mid \delta + |S_i| - 1, \Phi_{S_i S_i})}$$

is the ratio of products of evaluations of the IW density from (2.1). It follows from properties of the IW distribution that the normalizing constant for the *HIW* distribution is:

$$h(G, \delta, \Phi) = \frac{\prod_{i=1}^m \left| \frac{\Phi_{C_i C_i}}{2} \right|^{(\delta + |C_i| - 1)/2} \Gamma_{|C_i|} \left(\frac{\delta + |C_i| - 1}{2} \right)^{-1}}{\prod_{i=2}^m \left| \frac{\Phi_{S_i S_i}}{2} \right|^{(\delta + |S_i| - 1)/2} \Gamma_{|S_i|} \left(\frac{\delta + |S_i| - 1}{2} \right)^{-1}}. \quad (2.2)$$

2.2 Model Specification

We will treat a particular Bayesian Gaussian graphical model that is similar to models in work such as Armstrong et al. (2009). The considered prior on graphs, $P(G)$, is supported on the set of decomposable graphs with the probabilities of graphs proportional to

$$d^{|E|} (1 - d)^{\binom{p}{2} - |E|}. \quad (2.3)$$

The parameter d controls the sparsity of the graph. Conditional on G and hyperparameters δ and Φ , we let the covariance matrix Σ follow a *HIW*(G, δ, Φ) distribution, which has the consistency property that the submatrix distribution $P(\Sigma_{C_i C_i} \mid G, \delta, \Phi)$ is the same for any graph G with clique C_i (Dawid and Lauritzen, 1993). Larger δ makes the posterior more concentrated around the hyperparameter Φ . As in Carvalho and Scott (2009) and Donnet and Marin (2010), we choose $\delta = 1$. We use matrix hyperparameter $\Phi = c\mathcal{I}_p$, a scalar multiple of the $p \times p$ identity matrix. Larger c leads to larger graphs; see Jones et al. (2005).

Finally, suppose we observe a sample of n independent $\mathcal{N}_p(\mathbf{0}, \Sigma)$ random vectors $\mathbf{Y}_1, \dots, \mathbf{Y}_n$. Let \mathbf{Y} be the matrix with the vectors \mathbf{Y}_i as rows. The joint distribution of (\mathbf{Y}, G, Σ) then factors as

$$P(\mathbf{Y}, G, \Sigma \mid \delta, \Phi) = P(G)P(\Sigma \mid G, \delta, \Phi) \prod_{i=1}^n P(\mathbf{Y}_i \mid \Sigma).$$

Centering Gaussian data by subtracting off the sample mean and assuming mean $\mu = \mathbf{0}$ is standard practice since the distribution theory is essentially unchanged.

2.3 Metropolis-Hastings Sampler

We now briefly review the posterior sampling scheme used in prior work such as Armstrong et al. (2009). Define the sample covariance matrix

$$\mathbf{S} = (s_{jk}) = \frac{1}{n} \sum_{i=1}^n \mathbf{Y}_i \mathbf{Y}_i^T,$$

and set $\Phi^* = \Phi + n\mathbf{S}$ and $\delta^* = \delta + n$. Then

$$(\Sigma \mid \mathbf{Y}, \delta, \Phi, G) \sim HIW(G, \delta^*, \Phi^*) \quad (2.4)$$

and

$$P(\mathbf{Y} \mid \delta, \Phi, G) = \frac{1}{(2\pi)^{np/2}} \cdot \frac{h(G, \delta, \Phi)}{h(G, \delta^*, \Phi^*)}; \quad (2.5)$$

see Dawid and Lauritzen (1993) and Giudici and Green (1999).

Let $G = (V, E)$ and $G' = (V, E')$ be two decomposable graphs on $V = \{1, \dots, p\}$. Suppose that $C_1, S_2, C_2, \dots, S_m, C_m$ is a perfect sequence of the cliques and separators of G and that $\{j, k\} \in E$. If G' is equal to G except that the edge $\{j, k\}$ is removed, then the following three properties hold; see e.g. Armstrong et al. (2009). First, the edge $\{j, k\}$ is in a single clique C_q of G . Second, we have either $j \notin S_q$ or $k \notin S_q$. Third, suppose $k \notin S_q$, and let $C_{q_1} = C_q \setminus \{k\}$, $C_{q_2} = C_q \setminus \{j\}$ and $S_{q_2} = C_q \setminus \{j, k\}$. Then $C_1, S_2, \dots, S_q, C_{q_1}, S_{q_2}, C_{q_2}, S_{q+1}, \dots, S_m, C_m$ is a perfect ordering of all cliques and separators of G' . Let $\delta_2 = \delta + |S_{q_2}|$ and $\delta_2^* = \delta^* + |S_{q_2}|$. The above three observations can be used to show that the ratio of marginal likelihoods $P(\mathbf{Y} \mid G)/P(\mathbf{Y} \mid G')$ is equal to

$$\frac{h(G, \delta, \Phi)h(G', \delta^*, \Phi^*)}{h(G, \delta^*, \Phi^*)h(G', \delta, \Phi)} = \frac{|\Phi_{ee|S_{q_2}}|^{\left(\frac{\delta_2+1}{2}\right)} |\Phi_{jj|S_{q_2}}^* \Phi_{kk|S_{q_2}}^*|^{\left(\frac{\delta_2^*}{2}\right)}}{|\Phi_{ee|S_{q_2}}^*|^{\left(\frac{\delta_2^*+1}{2}\right)} |\Phi_{jj|S_{q_2}} \Phi_{kk|S_{q_2}}|^{\left(\frac{\delta_2}{2}\right)}} \times \frac{\Gamma\left(\frac{\delta_2}{2}\right) \Gamma\left(\frac{\delta_2^*+1}{2}\right)}{\Gamma\left(\frac{\delta_2+1}{2}\right) \Gamma\left(\frac{\delta_2^*}{2}\right)}. \quad (2.6)$$

Here, $e = \{j, k\}$ and $\Phi_{ee|S_{q_2}} = \Phi_{ee} - \Phi_{eS_{q_2}}(\Phi_{S_{q_2}S_{q_2}})^{-1}\Phi_{S_{q_2}e}$; the conditional variances for j and k are defined similarly. The ratio in (2.6) allows one to create a Markov chain with the posterior distribution $P(G \mid \mathbf{Y})$ as the stationary distribution by applying a Metropolis-Hastings procedure that avoids sampling of Σ .

Algorithm 1 (Gaussian). *Starting with a decomposable graph G_0 , repeat the following two steps for $t = 0, 1, 2, \dots$:*

- (i) *Create a graph G' by randomly picking an edge to delete from G_t or to add to G_t .*
- (ii) *If G' is decomposable, accept the move $G_{t+1} = G'$ with probability*

$$\min \left\{ 1, \frac{P(\mathbf{Y} \mid G')}{P(\mathbf{Y} \mid G)} \right\}, \quad (2.7)$$

setting $G_{t+1} = G$ if the move is rejected or G' is not decomposable.

Decomposability of the input G_0 can be tested with the Max-Cardinality algorithm (Cowell et al., 1999). Given the decomposable graph G_0 , the set of all decomposable graphs can be traversed following simple rules for edge addition and deletion (Giudici and Green, 1999).

3 Bayesian Graphical Models for t -Distributions

3.1 Classical and Alternative Multivariate t -Distributions

The *classical* multivariate t -distribution $t_{p,\nu}(\boldsymbol{\mu}, \boldsymbol{\Psi})$ in \mathbb{R}^p has density

$$f_\nu(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Psi}) = \frac{\Gamma(\frac{\nu+p}{2})|\boldsymbol{\Psi}|^{-1/2}}{(\pi\nu)^{p/2}\Gamma(\frac{\nu}{2})[1 + \delta_y(\boldsymbol{\mu}, \boldsymbol{\Psi})/\nu]^{(\nu+p)/2}}, \quad (3.1)$$

where $\delta_y(\boldsymbol{\mu}, \boldsymbol{\Psi}) = (\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ and $\mathbf{y} \in \mathbb{R}^p$. The vector $\boldsymbol{\mu} \in \mathbb{R}^p$ is the mean vector. The scale parameter matrix $\boldsymbol{\Psi} = (\psi_{jk})$ is assumed positive definite. For degrees of freedom $\nu > 2$, the covariance matrix exists and is equal to $\nu/(\nu - 2)$ times $\boldsymbol{\Psi}$. If $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Psi})$ is a multivariate normal random vector independent of the Gamma-random variable $\tau \sim \Gamma(\nu/2, \nu/2)$, then $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{X}/\sqrt{\tau}$ has a $t_{p,\nu}(\boldsymbol{\mu}, \boldsymbol{\Psi})$ -distribution (Kotz and Nadarajah, 2004, Chap. 1). The heavy tails of the distribution are related to small values of the divisor τ .

The classical t -distribution is useful for robust inference when only a handful of the observations are contaminated. When the dimension p is large, however, it is not uncommon for contamination to affect rather small parts of many observations. To handle such a situation better we consider p independent divisors $\tau_1, \dots, \tau_p \sim \Gamma(\nu/2, \nu/2)$. Assuming the divisors to be also independent of $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Psi})$, we create the random vector \mathbf{Y} with coordinates $Y_j = \mu_j + X_j/\sqrt{\tau_j}$ and define the *alternative* multivariate t -distribution to be the joint distribution of \mathbf{Y} . In symbols, $\mathbf{Y} \sim t_{p,\nu}^*(\boldsymbol{\mu}, \boldsymbol{\Psi})$. For robustified inference, the alternative t -distribution is appealing as it allows different rescaling of the different components of \mathbf{Y} .

3.2 Bayesian Inference With Classical t -Distributions

Suppose $\mathbf{Y}_1, \dots, \mathbf{Y}_n \in \mathbb{R}^p$ are independent random vectors drawn from the classical multivariate t -distribution $t_{\nu,p}(\boldsymbol{\mu}, \boldsymbol{\Psi})$. Let \mathbf{Y} be the matrix with the vectors \mathbf{Y}_i as rows. We are interested in the posterior distribution on graphs, $P(G \mid \mathbf{Y})$, where the graph G corresponds to conditional independence constraints on the latent multivariate normal vectors \mathbf{X}_i , that is, an off-diagonal entry θ_{jk} of the matrix $\boldsymbol{\Theta} = \boldsymbol{\Psi}^{-1}$ is zero unless $\{j, k\}$ is an edge in G .

In the normal model we can center the data by subtracting off the sample mean and assume, without loss of generality, that $\boldsymbol{\mu} = \mathbf{0}$. For the t -model, robust estimation of both $\boldsymbol{\mu}$ and $\boldsymbol{\Psi}$ is desirable, and we thus include $\boldsymbol{\mu}$ in our setup. Let $\boldsymbol{\tau} = (\tau_1, \dots, \tau_n)$ be the vector of unobserved Gamma-divisors for the n observations \mathbf{Y}_i . Proceeding as in the normal case, our full model factors the joint distribution of $\mathbf{Y}, \boldsymbol{\tau}, G, \boldsymbol{\Psi}, \boldsymbol{\mu}$ as

$$P(\mathbf{Y}, \boldsymbol{\tau}, G, \boldsymbol{\Psi}, \boldsymbol{\mu}) = P(G)P(\boldsymbol{\mu})P(\boldsymbol{\Psi} \mid G, \delta, \boldsymbol{\Phi}) \prod_{i=1}^n P(\mathbf{Y}_i \mid \tau_i, \boldsymbol{\Psi}, \boldsymbol{\mu})P(\tau_i \mid \nu), \quad (3.2)$$

where

$$(\mathbf{Y}_i \mid \tau_i, \boldsymbol{\Psi}) \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Psi}/\tau_i), \quad (\boldsymbol{\Psi} \mid G, \delta, \boldsymbol{\Phi}) \sim HIW(G, \delta, \boldsymbol{\Phi}), \quad (3.3)$$

$$(\tau_i \mid \nu) \sim \Gamma(\nu/2, \nu/2), \quad \boldsymbol{\mu} \sim \mathcal{N}_p(\mathbf{0}, \sigma_\mu \cdot \boldsymbol{\mathcal{I}}_p). \quad (3.4)$$

In later numerical work, we choose σ_μ large enough for the prior to be “flat” over a range of plausible values. Throughout the hyperparameters δ , Φ , and ν are fixed; recall Section 2.2.

Inference for the Gaussian case is simplified by integrating out the covariance matrix Σ ; recall (2.6). In the t -model, we may condition on τ and add/remove edges based on the ratio

$$\frac{P(\mathbf{Y} \mid G', \tau, \boldsymbol{\mu})}{P(\mathbf{Y} \mid G, \tau, \boldsymbol{\mu})} = \frac{h(G, \delta, \Phi)h(G', \delta^*, \Phi_\tau^*)}{h(G, \delta^*, \Phi_\tau^*)h(G', \delta, \Phi)}, \quad (3.5)$$

where

$$\mathbf{S}_{\tau \mathbf{Y} \mathbf{Y}}(\boldsymbol{\mu}) = \frac{1}{n} \sum_{i=1}^n \tau_i (\mathbf{Y}_i - \boldsymbol{\mu})(\mathbf{Y}_i - \boldsymbol{\mu})^T \quad \text{and} \quad \Phi_\tau^* = \Phi + n \mathbf{S}_{\tau \mathbf{Y} \mathbf{Y}}(\boldsymbol{\mu}). \quad (3.6)$$

Drawing τ given $G, \boldsymbol{\mu}$ and \mathbf{Y} is difficult, however, and we resort to conditioning on further parameters and consider the conditional distribution

$$(\tau_i \mid \mathbf{Y}, \boldsymbol{\mu}, \Psi) \sim \Gamma\left(\frac{\nu + p}{2}, \frac{\nu + \delta_{\mathbf{Y}_i}(\boldsymbol{\mu}, \Psi)}{2}\right); \quad (3.7)$$

compare Liu and Rubin (1995). This requires $\Theta = \Psi^{-1}$, and we use the method of Carvalho et al. (2007) to draw from

$$(\Psi \mid \mathbf{Y}, G, \tau, \boldsymbol{\mu}) \sim HIW(G, \delta^*, \Phi_\tau^*). \quad (3.8)$$

This procedure first draws $\Psi_{C_1 C_1}$, cycles through the cliques of G to draw $\Psi_{C_i C_i}$ given Ψ_{S_i, S_i} , and then uses a standard completion algorithm to determine the values of Ψ not associated with any clique. We then invert Ψ to obtain Θ . For decomposable graphs, the inversion can be done efficiently using the procedure of Dawid and Lauritzen (1993). That is, we compute

$$\Theta = \sum_{i=1}^m (\Psi_{C_i C_i})^{-1[0]} - \sum_{i=2}^m (\Psi_{S_i S_i})^{-1[0]}, \quad (3.9)$$

where $(\Psi_{C_i C_i})^{-1[0]}$ means that we take the $p \times p$ matrix of zeros and add in the elements of $(\Psi_{C_i C_i})^{-1}$ in their appropriate places. This calculation only requires the elements ψ_{jk} of Ψ corresponding to edges $\{j, k\} \in E$. Therefore, for the purposes of obtaining Θ , we need not perform the completion step in the method of Carvalho et al. (2007). Moreover, every step in the generation and inversion of Ψ is based on local computations at the clique level.

Now the conditional distribution $(\boldsymbol{\mu} \mid \mathbf{Y}, \tau, \Theta, G)$ is the multivariate normal distribution

$$\mathcal{N}_p \left(\left[\left(\sum_{i=1}^n \tau_i \right) \Theta + \Theta_\mu \right]^{-1} \Theta \left(\sum_{i=1}^n \tau_i \mathbf{Y}_i \right), \left[\left(\sum_{i=1}^n \tau_i \right) \Theta + \Theta_\mu \right]^{-1} \right) \quad (3.10)$$

where $\Theta_\mu = \mathcal{I}_p / \sigma_\mu$. To draw $\boldsymbol{\mu}$ using this conditional distribution we must invert the $p \times p$ matrix $(\sum_{i=1}^n \tau_i) \Theta + \Theta_\mu$, a potentially computationally expensive procedure. For practical applications, we thus simply set

$$\boldsymbol{\mu} = \frac{\sum_{i=1}^n \tau_i \mathbf{Y}_i}{\sum_{i=1}^n \tau_i} \quad (3.11)$$

instead of drawing from the conditional distribution in (3.10). We provide theoretical and numerical justifications for this alternative in the Appendix.

Algorithm 2 (Classical t). *Starting with a decomposable graph G_0 , and initial values $\boldsymbol{\mu}_0$ and $\boldsymbol{\tau}_0$, repeat the following steps for $t = 0, 1, 2, \dots$:*

(i) *Jointly draw a new graph G_{t+1} and a new matrix $\boldsymbol{\Theta}_{t+1}$ as follows:*

(a) *Draw G_{t+1} as in Algorithm 1, but using the ratio in (3.5).*

(b) *Conditional on $(\mathbf{Y}, G_{t+1}, \boldsymbol{\tau}_t, \boldsymbol{\mu}_t)$, sample $\boldsymbol{\Theta}_{t+1}$ by drawing $\boldsymbol{\Psi}_{t+1}$ from (3.8) and inverting it using (3.9).*

(ii) *Conditional on $(\mathbf{Y}, G_{t+1}, \boldsymbol{\Theta}_{t+1})$, sample the new independent components of the vector $\boldsymbol{\tau}_{t+1} = (\tau_{t+1,1}, \dots, \tau_{t+1,n})$ from (3.7).*

(iii) *Set $\boldsymbol{\mu}_{t+1}$ to the value in (3.11).*

In practice we hope to improve on the estimate of $P(G \mid \mathbf{Y})$ that we would obtain from the normal model. If we start with a “good” estimate of $\boldsymbol{\tau}$ as given by the *lasso* of Finegold and Drton (2011), we may be able to make considerable improvement over the normal model without sampling $\boldsymbol{\tau}$ after every edge draw. In our later simulations, we thus draw $\boldsymbol{\tau}$ only every $k > 1$ draws of $(G, \boldsymbol{\Theta})$, which does not affect the validity of the sampler. Note also that the $k - 1$ intermediate iterations do not involve $\boldsymbol{\Psi}$ (or its inverse $\boldsymbol{\Theta}$).

3.3 Bayesian Inference With Alternative t -Distributions

For the alternative t -model, there are only a few differences. The model itself is the same as in (3.2) except that now $\boldsymbol{\tau}_i$ is a p -vector, and

$$\begin{aligned} (\mathbf{Y}_i \mid \boldsymbol{\tau}_i, \boldsymbol{\Psi}, \boldsymbol{\mu}) &\sim \mathcal{N}_p(\boldsymbol{\mu}, \text{diag}(1/\sqrt{\boldsymbol{\tau}_i}) \cdot \boldsymbol{\Psi} \cdot \text{diag}(1/\sqrt{\boldsymbol{\tau}_i})), \\ (\tau_{ij} \mid \nu) &\sim \Gamma(\nu/2, \nu/2), \quad j = 1, \dots, p, \end{aligned}$$

with $\tau_{i1}, \dots, \tau_{ip}$ independent given ν . For the Gibbs sampler, we cannot draw the matrix $\boldsymbol{\tau} = (\tau_{ij})$ given $\mathbf{Y}, \boldsymbol{\Theta}, G, \boldsymbol{\mu}$ directly, but we can draw τ_{ij} given $\boldsymbol{\tau}_{i \setminus j}, \mathbf{Y}, \boldsymbol{\Theta}, G, \boldsymbol{\mu}$. The conditional density (derived in the Appendix) is

$$f(\tau_{ij} \mid \boldsymbol{\tau}_{i \setminus j}, \mathbf{Y}) = C(\alpha, \beta, \gamma) \cdot \tau_{ij}^{\alpha-1} \exp \{ -\tau_{ij}\beta - \sqrt{\tau_{ij}}\gamma \} \quad (3.12)$$

with

$$\alpha = \frac{\nu + 1}{2}, \quad \beta = \frac{\nu + (Y_{ij} - \mu_j)^2 \theta_{jj}}{2}, \quad \gamma = (Y_{ij} - \mu_j) \boldsymbol{\Theta}_{j \setminus j} \mathbf{X}_{i \setminus j},$$

and normalizing constant $C(\alpha, \beta, \gamma)$. The problem of sampling from this density also arose in the work of Finegold and Drton (2011), where the density appears in the context of a Markov chain Monte Carlo EM algorithm. We employ a rejection sampling scheme from Liu et al. (2012). This leads to the following Gibbs sampler.

Algorithm 3 (Alternative t). *Starting with a decomposable graph G_0 , and initial values $\boldsymbol{\mu}_0$ and $\boldsymbol{\tau}_0$, repeat the following steps for $t = 0, 1, 2, \dots$:*

- (i) *Jointly draw a new graph G_{t+1} and a new matrix $\boldsymbol{\Theta}_{t+1}$ as in Algorithm 2.*
- (ii) *For each observation $i = 1, \dots, n$, cycle through the variables $j = 1, \dots, p$ and draw τ_{ij} from its current full conditional in (3.12) to obtain a new matrix $\boldsymbol{\tau}_{t+1}$.*
- (iii) *Set $\boldsymbol{\mu}_{t+1}$ to the value in (3.11), where $\boldsymbol{\tau}$ is now a vector and the multiplications and divisions are done component-wise.*

This sampling scheme for the alternative model works well for moderate p ($p \approx 100$) and underlies our later simulations. The scheme becomes very computationally intensive, however, for large p , both in terms of the time to complete one iteration of step (ii) above and the number of iterations required to approach convergence of the Markov chain. It is conceivable that other strategies, such as using a Metropolis-Hastings step to sample from $P(\boldsymbol{\tau}|G, \mathbf{Y})$ directly, might perform better. However, we will not treat such alternative sampling schemes in the remainder of this paper, which is instead devoted to other models.

4 Dirichlet t -Models

We are faced with a trade-off between the classical and alternative models. If our goal is to identify pockets of contamination spread throughout a large data set, we certainly do not want to weight an entire observation via a single divisor as in the classical model. In the other extreme, with a different divisor for each variable, the alternative model proves to be computationally burdensome. Moreover, the alternative model has pairwise correlations bounded at a level that is somewhat restrictive for small degrees of freedom ν ; see Finegold and Drton (2011). The approach we propose in this section interpolates between the two extremes and seems appealing in particular when there are batches of variables taking on extreme values, while the rest exhibit behavior consistent with a normal model.

If groups of variables are similarly contaminated (or otherwise extreme), we can share statistical strength and ease our computational burden by forming clusters of Gamma divisors for each observation. We solve this clustering problem via a Dirichlet Process (DP) prior on the vector of $\boldsymbol{\tau}$ divisors for each observation. This Bayesian nonparametric approach avoids fixing a number of clusters and truly interpolates between the classical and alternative case.

4.1 Background on Dirichlet Processes

The Dirichlet Process is a measure on measures introduced by Ferguson (1973). Let P_0 be a probability measure on a measurable space $(\boldsymbol{\Theta}, \mathcal{B})$, and $\alpha > 0$. We say that P is distributed according to a Dirichlet process with parameters α and P_0 if for any finite measurable partition (A_1, \dots, A_r) of $\boldsymbol{\Theta}$, the random vector $(P(A_1), \dots, P(A_r))$ follows a Dirichlet distribution with parameters $(\alpha P_0(A_1), \dots, \alpha P_0(A_r))$. We write $P \sim DP(\alpha, P_0)$.

The Dirichlet process possesses a clustering property due to the fact that if $P \sim DP(\alpha, P_0)$ then P is discrete with probability 1. This holds even if the base measure P_0 is continuous (Ferguson, 1973). Let π_1, \dots, π_n be independent draws from a random measure $P \sim DP(\alpha, P_0)$.

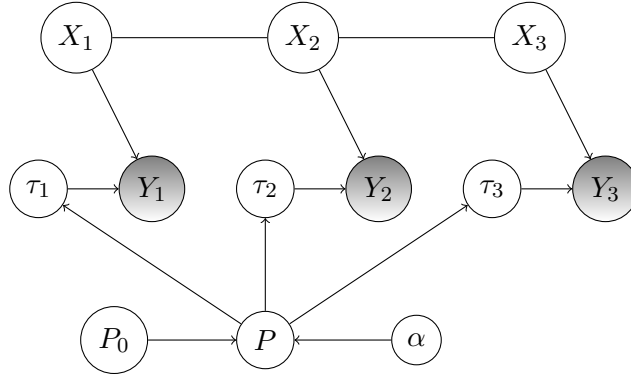


Figure 1: Representation of the process generating a t^α -random vector \mathbf{Y} from a latent normal random vector \mathbf{X} and an independent Dirichlet Gamma random vector $\boldsymbol{\tau}$. The missing undirected edge between X_1 and X_3 indicates a graphical conditional independence. Directed arrows illustrate the functional relationship among \mathbf{X} , $\boldsymbol{\tau}$, and \mathbf{Y} .

Then the conditional distribution of π_n given $\pi_{\setminus n}$ is a mixture, namely,

$$(\pi_n \mid \pi_{\setminus n}) \sim \frac{\alpha}{\alpha + n - 1} P_0(\pi_n) + \frac{n - 1}{\alpha + n - 1} \sum_{j=1}^{n-1} \delta_{\pi_j}(\pi_n), \quad (4.1)$$

where δ_{π_j} denotes a point mass at π_j . Hence, each new draw has a positive probability of assuming the same value as a previous draw, and this probability increases with each new draw. The choice of α greatly influences the number of expected clusters, with larger values leading to more clusters. New observations taking on the same values as existing ones in (4.1) gives an intuitive explanation to the phenomenon that Dirichlet processes often produce a small number of large clusters. This can be unsuitable for generic clustering applications but is, in fact, appealing for the robustification problem we consider. Here, we might often expect one large cluster that corresponds to uncontaminated (high-quality) observations.

4.2 Dirichlet t -Model

Applying Dirichlet processes in the t -distribution context yields the following construction, illustrated in Figure 1.

Definition 1. Let P_0 be the $\Gamma(\nu/2, \nu/2)$ distribution and let $P \sim DP(\alpha, P_0)$. For $j = 1, \dots, p$, let $\tau_j \sim P$ be independent of each other given P . We then say that the random vector $\boldsymbol{\tau} \in \mathbb{R}^p$ follows a Dirichlet Gamma distribution; in symbols $\boldsymbol{\tau} \sim D\Gamma_p(\alpha, \nu)$. If the random vectors $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Psi})$ and $\boldsymbol{\tau} \sim D\Gamma_p(\alpha, \nu)$ are independent, then we say that $\mathbf{Y} \in \mathbb{R}^p$ with coordinates $Y_j = \mu_j + X_j/\sqrt{\tau_j}$ follows a Dirichlet t -distribution; in symbols $\mathbf{Y} \sim t_{p,\nu}^\alpha(\boldsymbol{\mu}, \boldsymbol{\Psi})$.

The family of Dirichlet t -distribution includes the previous two models as extreme cases. When $\alpha \rightarrow 0$, we will have one cluster, giving us the classical t -distribution. When $\alpha \rightarrow \infty$ we will have p clusters, giving us the alternative t -distribution.

Unlike in the alternative model, there is no upper bound on the correlations between two variables *within* a cluster. For *any* two variables Y_j and Y_k , the marginal covariance is

$$\psi_{jk} \left[\frac{1}{\alpha+1} \frac{\nu}{\nu-2} + \frac{\alpha}{\alpha+1} \frac{\nu\Gamma((\nu-1)/2)^2}{2\Gamma(\nu/2)^2} \right] \quad (4.2)$$

where $1/(\alpha+1)$ is the probability that any two variables will be in the same τ cluster. As $\alpha \rightarrow 0$ we obtain a maximum covariance of $\psi_{jk} \cdot \nu/(\nu-2)$ and a maximum correlation of 1.

In contrast to the mixture model of Antoniak (1974), which uses n draws from a Dirichlet process for n observations, we here draw p values for the coordinates of a single observation. The relevant conditional distributions are similar to those given in Escobar and West (1995) but include terms that reflect that dependence among the p coordinates of an observation.

For Gibbs sampling we need the following full conditional (derived in the Appendix). For notational simplicity we consider \mathbf{Y} and $\boldsymbol{\tau}$ to be p -vectors representing a single observation:

$$(\tau_j \mid \boldsymbol{\tau}_{\setminus j}, \mathbf{Y}, \boldsymbol{\Theta}) \sim q_0 P_j(\tau_j) + \sum_{j' \neq j} q_{j'} \delta_{\tau_{j'}}(\tau_j). \quad (4.3)$$

This mixture distribution involves point masses, denoted $\delta_{\tau_k}(\tau_j)$ when supported at τ_k , and the distribution $P_j(\tau_j)$, which is the conditional distribution of τ_j given $(\boldsymbol{\tau}_{\setminus j}, \mathbf{Y}, \boldsymbol{\Theta})$ from the alternative t -model. The mixture weights for the point masses are denoted $q_{j'}$ and are proportional to the conditional density of $(Y_j \mid \boldsymbol{\Theta}, \mathbf{Y}_{\setminus j}, \boldsymbol{\tau}_{\setminus j}, \tau_j = \tau_{j'})$ evaluated at y_j . This density is that of a normal distribution with mean $\mu_c/\sqrt{\tau_{j'}}$ and variance $\sigma_c^2/\tau_{j'}$ evaluated at $y_j - \mu_j$. We denote the density as $f_{\mathcal{N}}(y_j - \mu_j; \mu_c/\sqrt{\tau_{j'}}, \sigma_c^2/\tau_{j'})$. Finally, the remaining mixture weight q_0 is proportional to α times the conditional density of $(Y_j \mid \boldsymbol{\Theta}, \mathbf{Y}_{\setminus j}, \boldsymbol{\tau}_{\setminus j})$ evaluated at y_j , where $\tau_j \sim P_0$. This density is that of a noncentral t -distribution with degrees of freedom ν and noncentrality parameter μ_c/σ_c , where μ_c and σ_c are the conditional mean and standard deviation of $(X_j \mid \mathbf{X}_{\setminus j})$. We denote the density as $f_T(y_j - \mu_j; \mu_c/\sigma_c, \nu)/\sigma_c$.

Now define a vector $\mathbf{z} \in \mathbb{R}^p$ of cluster indicators by setting $z_j = k$ if τ_j belongs to the k^{th} cluster. In our setting, all τ_j in the k^{th} cluster assume the same value, η_k , and thus $\boldsymbol{\tau}$ is a function of \mathbf{z} and the vector of cluster values $\boldsymbol{\eta}$. Hence, we may rewrite (4.3) as

$$(\tau_j \mid \boldsymbol{\tau}_{\setminus j}, \mathbf{Y}, \boldsymbol{\Theta}) \sim q_0 P_j(\tau_j) + \sum_{k=1}^K q_k \delta_{\eta_k}(\tau_j), \quad (4.4)$$

where K denotes the number of clusters, and q_k is proportional to

$$n_k^{(j)} \cdot f_{\mathcal{N}}(y_j - \mu_j; \mu_c/\sqrt{\eta_k}, \sigma_c^2/\eta_k)$$

with $n_k^{(j)} = |\{j' \neq j : z_{j'} = k\}|$. Rewriting (4.4) using the conditional cluster assignments gives

$$(z_j \mid \mathbf{z}_{\setminus j}, \boldsymbol{\eta}, \mathbf{Y}, \boldsymbol{\Theta}) \sim q_0 \delta_{z_{new}}(z_j) + \sum_{k=1}^K q_k \delta_k(z_j), \quad (4.5)$$

where $z_{new} = K + 1$ unless $n_{z_j}^{(j)} = 0$ in which case $z_{new} = z_j$.

The conditional in (4.5) describes the assignment of one node to a cluster given all the other cluster assignments and cluster values. We can also derive the conditional distribution of one cluster value given all the other cluster values and all the cluster assignments. Let $(k) := \{j : z_j = k\}$ and $n_k = |(k)|$. The conditional density (derived in the Appendix) is

$$f(\eta_k \mid \boldsymbol{\eta}_{\setminus k}, \mathbf{Y}, \boldsymbol{\Theta}, \mathbf{z}) = C(\alpha, \beta, \gamma) \cdot \eta_k^{\alpha-1} \exp\{-\eta_k \beta - \sqrt{\eta_k} \gamma\} \quad (4.6)$$

with

$$\alpha = \frac{\nu + n_k}{2}, \quad \beta = \frac{\nu + \text{tr}(\boldsymbol{\Theta}_{(k)(k)} \mathbf{Y}_{(k)} \mathbf{Y}_{(k)}^T)}{2}, \quad \gamma = \text{tr}(\boldsymbol{\Theta}_{(k)\setminus(k)} \mathbf{X}_{\setminus(k)} \mathbf{Y}_{(k)}^T).$$

The density being similar to (3.12), sampling can be performed using the same rejection sampling scheme. When the number of clusters is small relative to p , cycling through the clusters and drawing values for the whole cluster is much faster than cycling through all p nodes and assigning each to a cluster (and drawing values when new clusters are formed).

In the algorithm we may repeat some steps more frequently than others. For instance, if we are able to quickly identify clusters representing significant deviations from normality, then we can perform the third step more frequently than the computationally expensive reclustering step. For initial values, we use the *lasso* of Finegold and Drton (2011) to estimate $\boldsymbol{\tau}$, which means we have one cluster for each observation.

4.3 Inference for the Clustering Parameter α

The Dirichlet Process parameter α plays a key role in determining the number of clusters, and it is beneficial to add another level of hierarchy and place a prior on α . Following Escobar and West (1995) who treat Dirichlet process mixture models, we consider a $\Gamma(a, b)$ prior on α . In practice we choose a and b to give low prior mean to α , which leads to fewer clusters and easier computation, but allows for more clusters as the data requires.

Let k denote the number of clusters. We have from Antoniak (1974) that, for $k = 1, \dots, n$,

$$P(k \mid \alpha, n) = c_n(k) n! \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}, \quad (4.7)$$

where $c_n(k) = P(k \mid \alpha = 1, n)$. Posterior inference is simplified by the fact that α is conditionally independent of the observed data given the cluster assignments, leading to

$$P(\alpha \mid k, \pi, \mathbf{X}) = P(\alpha \mid k) \propto P(\alpha) P(k \mid \alpha). \quad (4.8)$$

That is, inference is based on the prior for α and a single observation from $P(k \mid \alpha, n)$.

Returning to our setting with an n -sample, let the vector $\mathbf{k} = (k_1, \dots, k_n)$ comprise the numbers of clusters in the n observations. Then α is conditionally independent of $(\boldsymbol{\tau}, \mathbf{Y}, \boldsymbol{\Theta}, \boldsymbol{\mu}, G)$ given \mathbf{k} , and

$$P(\alpha \mid \mathbf{k}, \boldsymbol{\tau}, \mathbf{Y}, \boldsymbol{\Theta}, \boldsymbol{\mu}, G) \propto P(\alpha) P(\mathbf{k} \mid \alpha) = P(\alpha) \prod_{i=1}^n P(k_i \mid \alpha).$$

Let $\beta(\alpha, \beta)$ be the Beta function. Generalizing the results of Escobar and West (1995) to multiple observations k_i , we use the fact that

$$\frac{\Gamma(\alpha)}{\Gamma(\alpha + p)} = \frac{(\alpha + n)\beta(\alpha + 1, p)}{\alpha\Gamma(p)} \quad (4.9)$$

to obtain that $P(\alpha | \mathbf{k})$ is proportional to

$$P(\alpha) \propto \sum_{k_i=n} (\alpha + p)^n \int_0^1 w_1^\alpha (1 - w_1)^{p-1} dw_1 \times \cdots \times \int_0^1 w_n^\alpha (1 - w_n)^{p-1} dw_n. \quad (4.10)$$

Hence, we may view $P(\alpha | \mathbf{k})$ as a marginal distribution of $P(\alpha, w_1, \dots, w_n | \mathbf{k})$ where $0 < w_i < 1$ are random variables that are conditionally independent of each other given α . Writing $\mathbf{w} = (w_1, \dots, w_n)$, we consider the conditional distribution

$$P(\alpha | \mathbf{w}, \mathbf{k}) \propto \alpha^{a + \sum_{i=1}^n k_i - n - 1} (\alpha + p)^n \exp \left\{ -\alpha \left(b - \sum_{i=1}^n \log w_i \right) \right\}.$$

Expanding $(\alpha + p)^n$ gives a mixture of $n + 1$ Gamma-distributions, namely,

$$(\alpha | \mathbf{w}, \mathbf{k}) \sim \sum_{j=0}^n \pi_j \Gamma \left(a + \sum_{i=1}^n k_i - j, b - \sum_{i=1}^n \log w_i \right) \quad (4.11)$$

with

$$\pi_j \propto \binom{n}{j} p^j \left(b - \sum_{i=1}^n \log w_i \right)^j \Gamma \left(a + \sum_{i=1}^n k_i - j \right). \quad (4.12)$$

The w_i being conditionally independent given α and \mathbf{k} , it holds that

$$(w_i | \alpha, \mathbf{k}, \mathbf{w}_{\setminus i}) \sim \beta(\alpha + 1, p). \quad (4.13)$$

Now suppose we observe n independent random vectors $\mathbf{Y}_1, \dots, \mathbf{Y}_n \in \mathbb{R}^p$ that follow a $t_{p,\nu}^\alpha(\boldsymbol{\mu}, \boldsymbol{\Psi})$ distribution. Once again, let \mathbf{Y} and $\boldsymbol{\tau}$ be the associated matrices of observations and divisors. For small α , i.e., few expected clusters, we can create a sampler as follows. Let k_i be the number of clusters for the i^{th} observation. The state space consists of the values for $(G, \boldsymbol{\Theta}, \mathbf{z}, \boldsymbol{\eta})$ where $\mathbf{z} = \{z_{ij}\}$ is now an $n \times p$ matrix and $\boldsymbol{\eta}$ an array collecting n vectors of length k_1, \dots, k_n . Following Teh et al. (2006), we propose the following Gibbs sampler.

Algorithm 4 (Dirichlet t). *Starting with a decomposable graph G_0 , and initial values $\boldsymbol{\mu}_0, \mathbf{z}_0, \alpha_0$, and $\boldsymbol{\eta}_0$, repeat the following steps for $t = 0, 1, 2, \dots$:*

- (i) *Jointly draw a new graph G_{t+1} and a new matrix $\boldsymbol{\Theta}_{t+1}$ as in Algorithm 2.*
- (ii) *For each observation $i = 1, \dots, n$, cycle through the variables $j = 1, \dots, p$ and draw z_{ij} from the conditional given in (4.5). If $z_{ij} = z_{\text{new}}$ assign to this new cluster a value $\eta_{iz_{\text{new}}}$ by sampling from $P_j(\tau_j)$ in (4.3). This results in a new matrix \mathbf{z}_{t+1} .*
- (iii) *For each observation $i = 1, \dots, n$, cycle through the clusters $k = 1, \dots, K_i$ and draw η_{ik} using (4.6). This results in a new array $\boldsymbol{\eta}_{t+1}$.*
- (iv) *Assign $\boldsymbol{\mu}_{t+1}$ as in Algorithm 3.*
- (v) *For each observation $i = 1, \dots, n$, draw w_i from the conditional given in (4.13).*
- (vi) *Draw α from the conditional given in (4.11).*

5 Simulations

5.1 AR1 with $p=25$

To illustrate the behaviour of the different Bayesian methods, we first present simulations for graphs with 25 nodes, for which we run the Markov chain Monte Carlo samplers for 10,000 iterations per possible edge, as suggested in Jones et al. (2005). We choose a graph for an autoregressive process of order one, that is, the nodes form a chain and the corresponding precision matrix Θ is tri-diagonal. We forego simulating random draws of Θ for a clear distinction between true positives and negatives. Instead, we set the non-zero off-diagonal elements of Θ to -1 and the diagonal elements to 3 (except the first and last, which are set to 2). Unless otherwise noted, we fix the degrees of freedom $\nu = 3$, the graph prior parameter $d = 0.05$, recall (2.3), and the hyperparameter $\delta = 1$. It is not desirable, however, to have exactly the same priors in the normal and t -models. With the most extreme observations downweighted, the entries of the matrix $\mathbf{S}_{\tau\mathbf{Y}\mathbf{Y}}$ from the t -model tend to be smaller in magnitude than those of the sample covariance matrix \mathbf{S} , and, in our experience, the same hyperparameter matrix Φ then leads to larger graphs for the t -case. To get graphs of comparable size, we choose $\Phi = \mathbf{I}_p/5$ for the normal model and $\Phi = \mathbf{I}_p/10$ for t -models.

We first simulate n independent normal observations from $\mathcal{N}_p(\mathbf{0}, \Theta^{-1})$. In order to illustrate the effects on inference of dependence structures, we assume the mean to be known and run five different estimation methods:

- (a) The normal procedure (§2).
- (b) The normal procedure using the maximum likelihood estimate $\mathbf{S}_{\hat{\tau}\mathbf{Y}\mathbf{Y}}$ from the classical $t_{p,3}$ -model as the sufficient statistic instead of \mathbf{S} .
- (c) The classical $t_{p,3}$ procedure (§3.2), drawing the matrix τ once every 10 edge proposals.
- (d) The Dirichlet $t_{p,3}^\alpha$ procedure (§4.2) with a $\Gamma(1,1)$ prior on α . We draw new cluster identifiers \mathbf{z} for every 20 draws of η , which in turn is drawn once every 10 edge proposals.
- (e) The alternative $t_{p,3}^*$ procedure (§3.3), drawing the matrix τ once every 10 edge proposals.

Assuming known means only favors normal procedures for which estimation of the mean from heavy-tailed data is problematic in itself. For each method, we perform 3 million edge proposals. We repeat the process 50 times, each time recording the posterior probability $P(e_{jk} = 1 \mid \mathbf{Y})$, that is, the probability of edge $\{j, k\}$ to be in the true edge set. If $P(e_{jk} = 1 \mid \mathbf{Y}) > \epsilon$ for some threshold ϵ , we consider it a “positive”. We let ϵ range from 0 to 1 and record the number of true and false positives in all 50 simulations. We then compare the true positive rate to the false positive rate at each threshold. This entire process is repeated for data generated from a $t_{p,3}(\mathbf{0}, \Theta^{-1})$ and a $t_{p,3}^*(\mathbf{0}, \Theta^{-1})$ distribution.

The simulation results are summarized in Figure 2, which shows that the more flexible models indeed perform better when the data is generated from the more complicated model. With normal data, the t -models all perform similarly well and the normal model outperforms

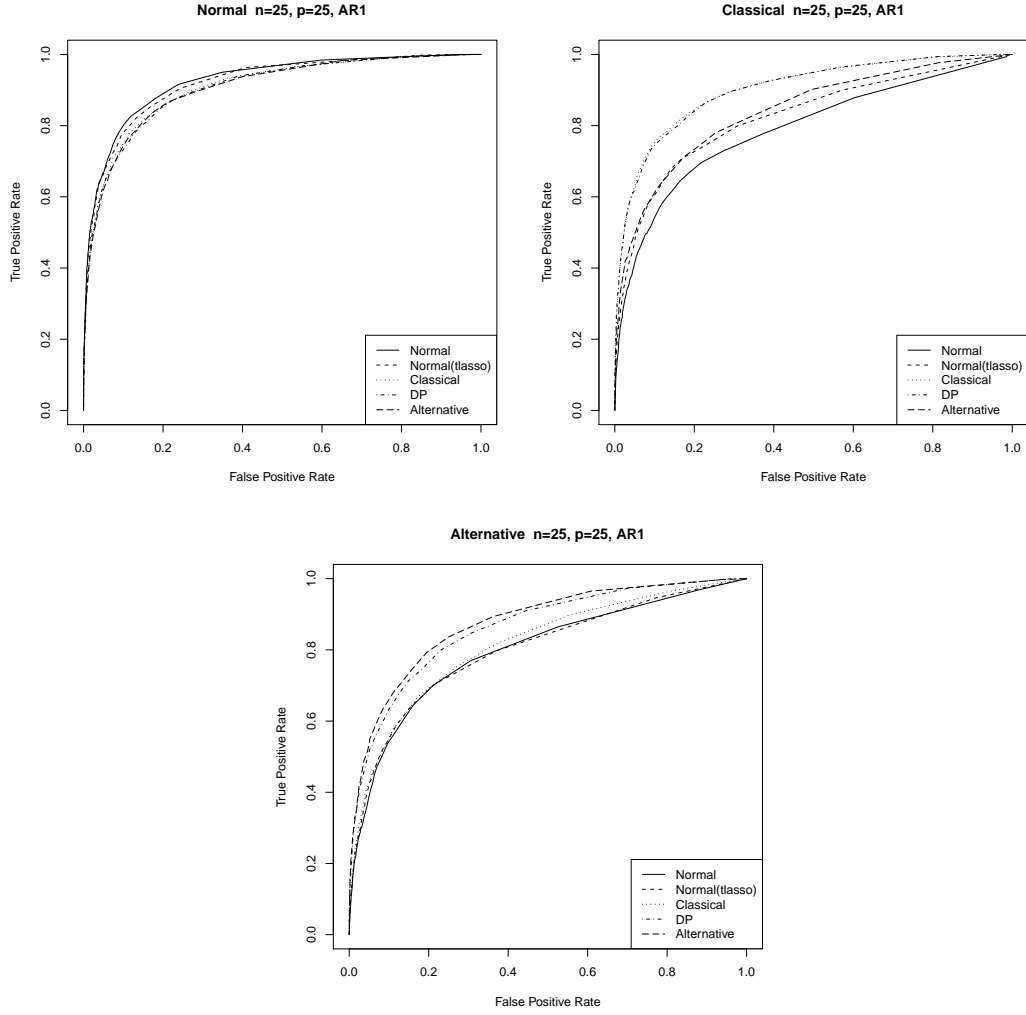


Figure 2: ROC curves depicting the performances of the five methods for data generated from a $\mathcal{N}_{25}(\mathbf{0}, \mathbf{\Theta}^{-1})$ distribution, a $t_{25,3}(\mathbf{0}, \mathbf{\Theta}^{-1})$ distribution and a $t_{25,3}^*(\mathbf{0}, \mathbf{\Theta}^{-1})$ distribution.

them only slightly. For classical t -data, the classical t -model performs significantly better than the normal model. The alternative model is clearly inferior to the classical model. The performance of the Dirichlet t -model, on the other hand, is barely distinguishable from that of the classical t -model. With a $\Gamma(1, 1)$ prior for α , the Dirichlet method finds an average of 1.4 clusters per τ vector, rendering it very similar to the classical model. The normal procedure with the robustified estimate $\mathbf{S}_{\hat{\tau}\mathbf{Y}\mathbf{Y}}$ performs better than the purely normal technique, but not as well as the classical t -model. In addition to the superior performance seen here, the fully Bayesian approach has, of course, the added benefit of a posterior distribution on τ . This can be useful to assess of each observation’s consistency with normality.

For alternative t -data, the alternative t -method clearly outperforms its normal and classical t -analogues, which perform equally poorly. Only the Dirichlet t -model, which finds an average of 7.7 clusters per τ vector, performs comparably to the alternative model. Its strong performance on both classical and alternative data suggests that the Dirichlet method is indeed an effective compromise between the two other t -distribution techniques.

For classical t -data, fitting the classical, the Dirichlet and the alternative t -model took on average $2.3N$, $3.6N$ and $4.5N$, respectively, where N is the processing time for the normal model. Based on use of ‘R’ (R Development Core Team, 2010), the times are meant only to be rough estimates of actual computational complexity. Nevertheless, the comparison suggests that the Dirichlet approach adaptively produces statistically efficient estimates while using a run-time about halfway between that of the classical and alternative procedures. For alternative t -data, the Dirichlet model faces the added complexity of reclustering steps without much benefit from any clustering. Indeed, the average run time was $5N$ for the Dirichlet model compared to $3.8N$ for the alternative model. This said, we would not expect any real application requiring as large a number of clusters as simulated alternative t -data.

5.2 Random Graphs with $p=100$

For a more challenging scenario, we consider $p = 100$ nodes and create the graph by forming 20 random cliques of size 2 to 5. For each clique, we pick nodes at random and form edges between all nodes in the clique. We draw the mean vector μ as a p -vector of independent standard normals. We set the non-zero entries in Θ as before (but multiply the diagonal entries by a constant to ensure a minimum eigenvalue of at least 0.6). We then simulate $n = 100$ independent observations from a $\mathcal{N}_p(\mathbf{0}, \Theta^{-1})$ distribution to create latent data X .

Next, we contaminate the data via an $n \times p$ matrix τ that holds divisors for X , with the goal of creating contamination in many parts of many observations so that detecting outliers manually would be difficult. For each one of a total of 10 contaminations, we draw a Poisson number of rows and a Poisson number of columns (mean 10 for both). We then select uniformly at random a submatrix of τ that has this given size and assign a single random value (uniform[0.01,0.2]) to the entries in the submatrix. The remaining entries of τ are set to 1. The observations Y are created by setting $Y_{ij} = \mu_j + X_{ij}/\tau_{ij}$. This results on average in contamination in slightly less than one in ten elements of the latent data matrix.

We run the five algorithms from Section 5.1 under the settings described there, but with

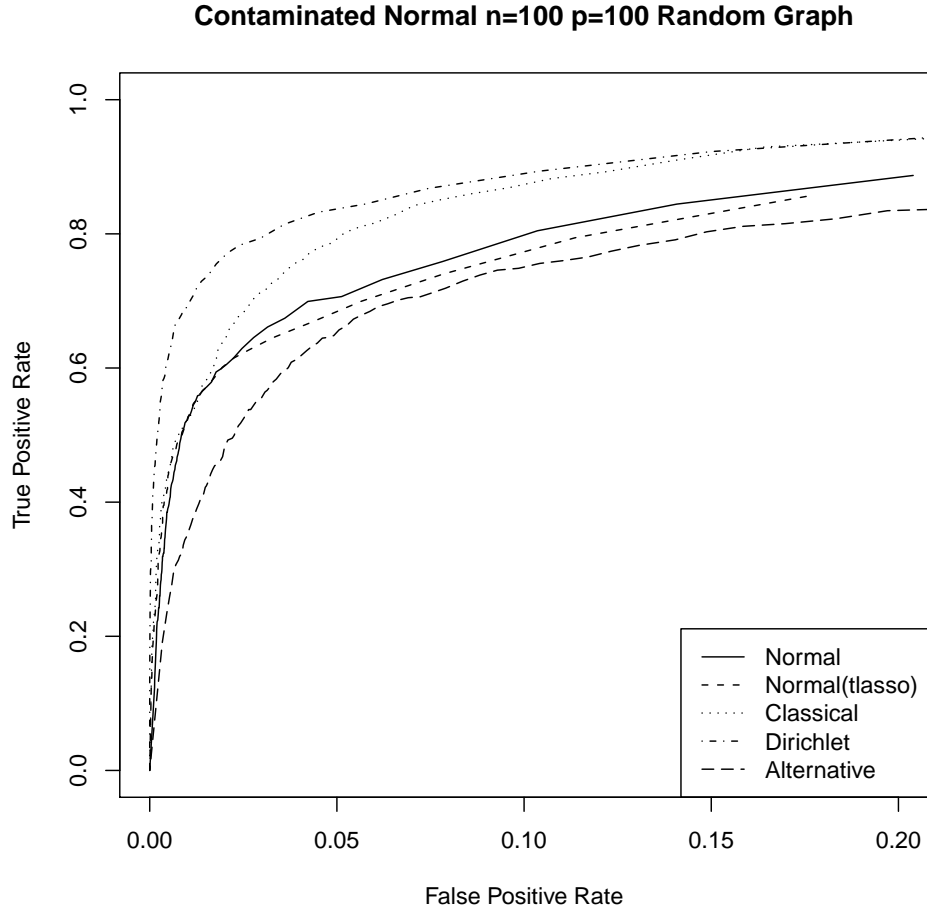


Figure 3: ROC curves depicting the performances of the five methods for data generated from a contaminated $\mathcal{N}_{100}(\mathbf{0}, \Theta^{-1})$ distribution that is Markov to a random graph.

$\Phi = \mathcal{I}_p/5$ for the normal model and $\Phi = \mathcal{I}_p/20$ for t -models, to get graphs of comparable size. We run the samplers to obtain 2 million edge draws and for the t models we sample τ every 30 edge draws. The results for 25 repeats of the entire process are shown in Figure 3, which makes a clear case for the Dirichlet t -model.

The 2 million draws do not seem enough for “convergence” of all Markov chains. Figure 4, which shows the number of edges in the estimated graph over the iterations of the samplers, suggests that the alternative model may require much longer runs. As another plus, the Dirichlet model seems to require far fewer iterations.

6 Gene Expression Data

Gasch et al. (2000) present data from microarray experiments with yeast strands. As in the

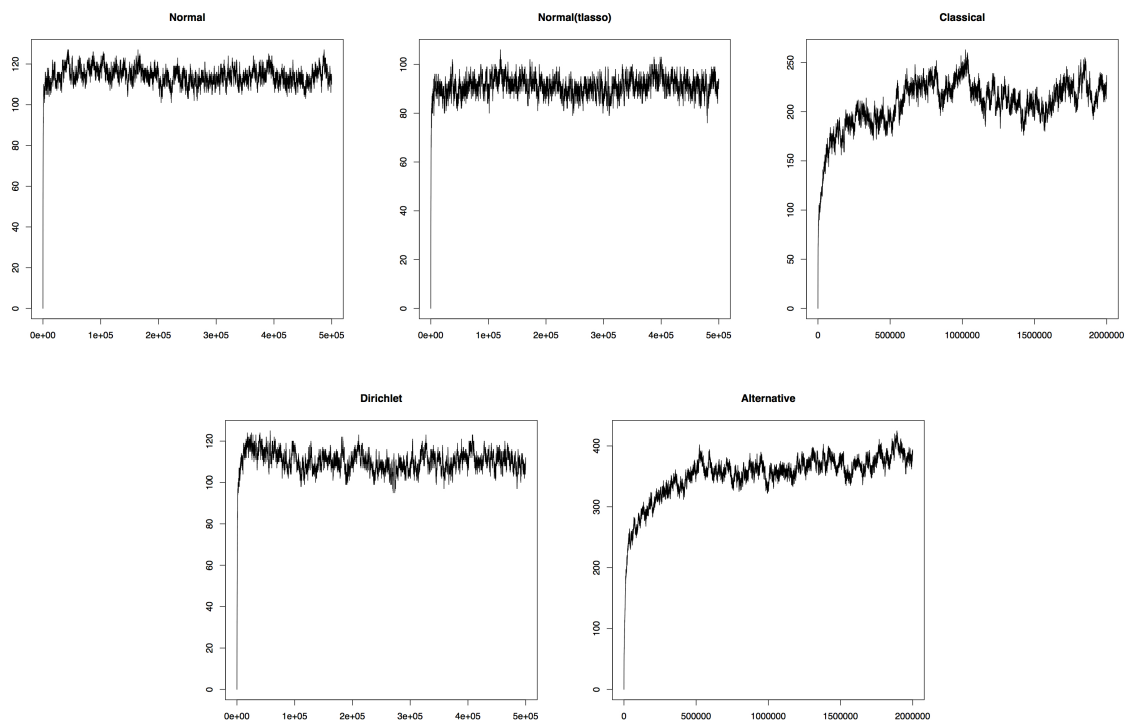


Figure 4: Edges in estimated graph plotted against iterations of the Gibbs sampler for models used in the contaminated normal simulations.

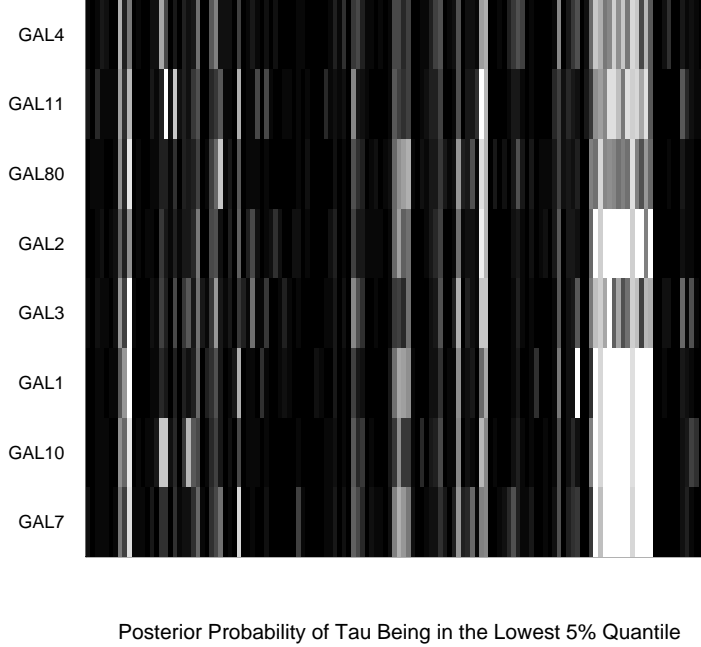


Figure 5: Lighter colors represent higher posterior probability that τ_{ij} is in the bottom 5% quantile of the prior distribution on τ . The light blocks on the bottom right correspond to the “outliers” in the yeast data from Section 6.

frequentist work in Finegold and Drton (2011), we focus on 8 genes involved in galactose utilization; 136 experiments have data for all the 8 genes. In 11 experiments, 4 of the genes have abnormally large negative expression values. Figure 5, which is obtained from 3 million iterations of our sampler, shows that the Dirichlet $t_{3,8}^{\alpha=1}$ model leads to downweighting of the abnormal values but not entire observations, as desired.

To demonstrate automatic detection of outliers in a much larger dataset, we add the data for 92 genes, selected at random from those without missing data. We fit the models $t_{3,100}^{\alpha}$ for fixed $\alpha = \{1, 10, 100\}$. We run the samplers for 2 million edge draws and follow a proposal of Donnet and Marin (2010) to select candidate edges with probability proportional to the sample correlation of the corresponding variables. Figure 6 provides convergence diagnostics.

For $\alpha = 100$ the propensity to cluster is very weak with an average of 67 different clusters in each row of τ in the final iteration. Nonetheless, the average τ value for the 4 relevant genes in the 11 relevant experiments was 0.046 compared with an average value of 0.84 for the rest of the matrix. That is, as we let the Dirichlet t model approach the alternative model, we achieve downweights of the suspected outliers even without the benefits of clustering.

With $\alpha = 10$ the propensity to cluster is a bit stronger; the rows of τ have an average of 20 different clusters in the final iteration. We achieve relative downweighting similar to the $\alpha = 100$ model (0.067 to 0.47). Despite 20 clusters per row, at least two of the outliers always

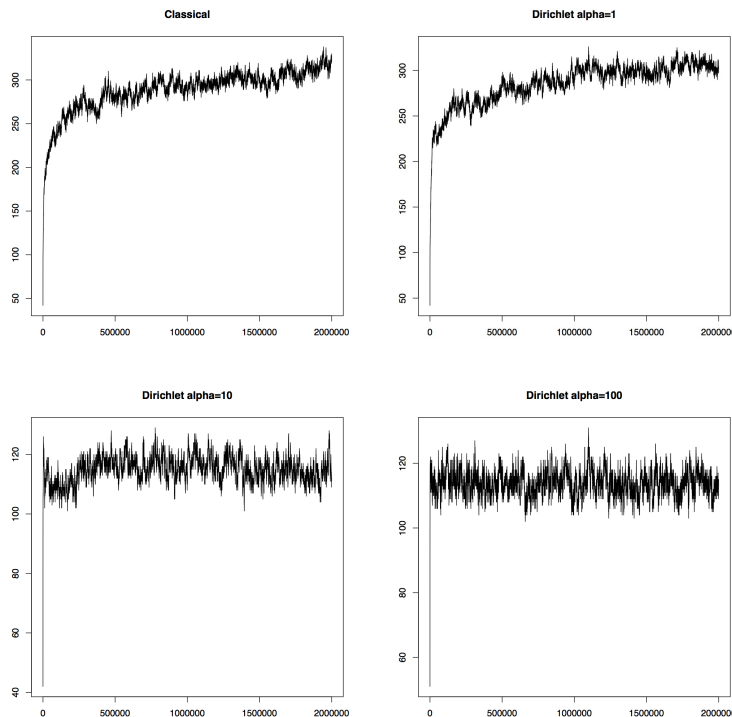


Figure 6: Number of edges in estimated graph plotted against iterations of the Gibbs sampler for models used in analysis of gene expression data.

share a cluster and in five of the experiments, at least three share a cluster.

Letting $\alpha = 1$, we achieve much greater clustering (2.8 per row). The suspected outliers tend to cluster together, but there is little practical difference between this model and the classical t . The classical t downweights most of the observations significantly and takes much longer to reach the appearance of equilibrium in the estimated edges; see Figure 6.

7 Discussion

We have extended Bayesian approaches to graphical Gaussian modeling using three variations of multivariate t -distributions. While these extensions all come at increased computational expense, they can have substantial statistical benefit. In particular, one obtains a posterior distribution for latent weights that measures uncertainty about potential outliers.

Our extensions to t -distributions are based on one particular Gaussian model, but many other variations have been treated in the literature. Some authors place a prior on the degrees of freedom parameter δ for the Hyper Inverse Wishart (HIW) prior distribution. In addition, one can introduce a prior on the edge density parameter d from (2.3). An alternative approach is to place a uniform prior on the size of the graph, and then a uniform prior on all graphs of the same size. We have chosen the scale matrix of the HIW prior to be $\Phi = c\mathcal{I}_p$ and

treated fixed choices of c in our simulations. One could instead include a prior on c , or set $\Phi = c\mathbf{A}$, where \mathbf{A} is the sample covariance matrix, or an equicorrelated matrix with diagonal elements equal to 1 and off-diagonal elements equal to a common value ρ . Armstrong et al. (2009) consider all of these variations, including placing a prior on ρ . Finally, as mentioned in the introduction, there now exist more flexible versions of the HIW distribution as well as techniques for approximate computations for non-decomposable graphs. Incorporating the former generalization in the t -distribution context would be straightforward; addressing non-decomposable graphs, however, constitutes an interesting area for further research.

As noted previously by many authors, for large p even the most likely graphs may have very small posterior probability, making it difficult and not necessarily very informative to identify the graph with highest posterior probability. In practice, the focus may thus often be on more modest goals, such as the posterior distribution of subgraphs on some subset of vertices, or even more simply, the marginal posterior probability that each edge is in the edge set E that we considered in the simulation study.

In the data from Section 6, a group of genes had extreme expression values in several observations. While the Dirichlet t -model did a good job identifying these clusters, it could potentially be worthwhile to share statistical strength by explicitly modeling the clustering of latent weights as similar across observations. This could be done by treating the p -vectors τ_1, \dots, τ_n as draws from a Dirichlet process mixture model. Combined with the Dirichlet t -model, this would give a ‘doubly Dirichlet’ τ matrix. That is, we will have $k \leq p$ distinct elements within each row (observation) and $l \leq n$ distinct rows. Inference in this model would be more involved than in the ordinary Dirichlet t -model we discussed, but the full conditionals necessary to devise a Gibbs sampler would be of similar type.

Acknowledgements

This work was supported by the U.S. National Science Foundation under Grant No. DMS-0746265. Mathias Drton was also supported by an Alfred P. Sloan Fellowship.

References

- Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems. *Annals of Statistics* **2**, 1152–1174.
- Armstrong, H., Carter, C. K., Wong, K. F., and Kohn, R. (2009). Bayesian covariance matrix estimation using a mixture of decomposable graphical models. *Statistics and Computing* **19**, 303–316.
- Atay-Kayis, A. and Massam, H. (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika* **92**, 317–335.
- Carvalho, C. and Scott, J. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika* **96**, 1–16.

- Carvalho, C. M., Massam, H., and West, M. (2007). Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika* **94**, 647–659.
- Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems*. Statistics for Engineering and Information Science. Springer-Verlag, New York.
- Dawid, A. P. and Lauritzen, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics* **21**, 1272–1317.
- De Finetti, B. (1961). The Bayesian approach to the rejection of outliers. In *Proceedings of the Fourth Berkeley Symposium on Probability and Statistics*, volume 1, pages 199–210.
- Dellaportas, P., Giudici, P., and Roberts, G. (2003). Bayesian inference for nondecomposable graphical Gaussian models. *Sankhyā* **65**, 43–55.
- Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis* **90**, 196–212.
- Donnet, S. and Marin, J.-M. (2010). An empirical Bayes procedure for the selection of Gaussian graphical models. arXiv:1003.5851.
- Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.
- Finegold, M. and Drton, M. (2011). Robust graphical modeling with classical and alternative t -distributions. *Annals of Applied Statistics* **5**, 1057–1080.
- Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell* **11**, 4241–4257.
- Giudici, P. and Green, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika* **86**, 785–801.
- Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science* **20**, 388–400.
- Kotz, S. and Nadarajah, S. (2004). *Multivariate t distributions and their applications*. Cambridge University Press, Cambridge.
- Lauritzen, S. L. (1996). *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York. Oxford Science Publications.

- Lenkoski, A. and Dobra, A. (2011). Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *Journal of Computational and Graphical Statistics* **20**, 140–157.
- Liu, C. and Rubin, D. B. (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica* **5**, 19–39.
- Liu, Y., Wichura, M. J., and Drton, M. (2012). Rejection sampling for an extended Gamma distribution. Manuscript.
- R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rajaratnam, B., Massam, H., and Carvalho, C. M. (2008). Flexible covariance estimation in graphical Gaussian models. *Annals of Statistics* **36**, 2818–2849.
- Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics* **29**, 391–411.
- Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101**, 1566–1581.
- West, M. (1984). Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society, Series B* **46**, 431–439.
- Yuan, M. and Huang, J. (2009). Regularized parameter estimation of high dimensional t distribution. *Journal of Statistical Planning and Inference* **139**, 2284–2292.

A Appendix

A.1 Alternative Procedures for Drawing the Mean Vector $\boldsymbol{\mu}$

The parameters of the normal distribution in (3.10) involve the inverse of the sum of matrices $\sum_{i=1}^n \tau_i \boldsymbol{\Theta} + \boldsymbol{\Theta}_\mu$, where $\boldsymbol{\Theta}_\mu = \mathcal{I}_p / \sigma_\mu$ is a multiple of the identity matrix. With rare exceptions, the hyperparameter σ_μ , a variance, is chosen large so as to make the prior distribution on $\boldsymbol{\mu}$ flat. In those cases, $\boldsymbol{\Theta}_\mu$ is small compared to the typical values of $\sum_{i=1}^n \tau_i \boldsymbol{\Theta}$. Hence, little is lost by ignoring the term $\boldsymbol{\Theta}_\mu$ in the matrix inversion, which leads to the distributional approximation

$$(\boldsymbol{\mu} \mid \mathbf{Y}, G, \boldsymbol{\tau}, \boldsymbol{\Theta}) \approx \mathcal{N}_p \left(\frac{1}{\sum_{i=1}^n \tau_i} \cdot \sum_{i=1}^n \tau_i \mathbf{Y}_i, \frac{1}{\sum_{i=1}^n \tau_i} \cdot \boldsymbol{\Theta}^{-1} \right). \quad (\text{A.1})$$

This approximation still requires the completion step we have avoided to obtain $\boldsymbol{\Theta}^{-1} = \boldsymbol{\Psi}$. While this is not prohibitively expensive, we find that simply setting $\boldsymbol{\mu}$ equal to the mean of the distribution in (A.1) works well enough in practice – we call this “Robust Centering”. To test this, we simulated classical t data from a chain graph with 25 nodes as described in Section 5. We ran four versions of the the Dirichlet t algorithm (Algorithm 4) starting with the same seed: using naive centering (subtract of the sample mean for each variable and set $\boldsymbol{\mu} = \mathbf{0}$); Robust Centering; sampling from the approximate conditional in equation (A.1); and sampling from the exact conditional in (3.10). With σ_μ set to 100,000 we find virtually no difference between the estimated values of $\boldsymbol{\mu}$ from the last three procedures, but significant difference between those three and the first. The first procedure does a worse job of estimating $\boldsymbol{\mu}$ and, as a result, $\boldsymbol{\Theta}$. We conclude that Robust Centering is better than naive centering, but that virtually nothing is lost by failing to sample from the approximate or full conditionals.

A.2 Full Conditional for Latent Divisors in the Dirichlet t -Model

For notational convenience, let $z_j = 0$ if τ_j belongs to a new cluster and consider the case $j = p$. Let K be the number of distinct clusters containing elements other than τ_p . We may then write

$$P(\tau_p \leq t \mid \boldsymbol{\tau}_{\setminus p}, \mathbf{Y}, \boldsymbol{\Theta}) = \sum_{k=0}^K P(\tau_p \leq t \mid \boldsymbol{\tau}_{\setminus p}, \mathbf{Y}, z_p = k) P(z_p = k \mid \boldsymbol{\tau}_{\setminus p}, \mathbf{Y}, \boldsymbol{\Theta}).$$

The conditional density of τ_p given $(\boldsymbol{\tau}_{\setminus p}, \mathbf{Y}, \boldsymbol{\Theta}, z = k)$ is trivially the point mass at η_k , the value assumed by all elements of $\boldsymbol{\tau}$ that belong to the k^{th} cluster. The conditional density of τ_p given $(\boldsymbol{\tau}_{\setminus p}, \mathbf{Y}, \boldsymbol{\Theta}, z = 0)$ is

$$\begin{aligned} f(\tau_p \mid \boldsymbol{\tau}_{\setminus p}, \mathbf{Y}, \boldsymbol{\Theta}, z = 0) &\propto f(\tau_p \mid \boldsymbol{\tau}_{\setminus p}, \boldsymbol{\Theta}, z = 0) f(\mathbf{Y} \mid \boldsymbol{\tau}, \boldsymbol{\Theta}, z = 0) \\ &= f(\tau_p \mid z = 0) f(\mathbf{Y} \mid \boldsymbol{\tau}, \boldsymbol{\Theta}) \\ &\propto f_\Gamma(\tau_p; \nu/2, \nu/2) f_\mathcal{N}(\mathbf{Y}_p; \mu_c / \sqrt{\tau_p}, \sigma_c^2 / \tau_p), \end{aligned} \quad (\text{A.2})$$

where $f_{\Gamma}(\tau_p; \nu/2, \nu/2)$ is the density of a $\Gamma(\nu/2, \nu/2)$ distribution evaluated at τ_p . The distribution specified by (A.2) is the Gibbs sampling distribution from the alternative model in (3.12). Now,

$$\begin{aligned} P(z_p = 0 \mid \tau_{\setminus p}, \mathbf{Y}, \boldsymbol{\Theta}) &\propto P(z_p = 0, Y_p \mid \mathbf{Y}_{\setminus p}, \tau_{\setminus p}, \boldsymbol{\Theta}) \\ &= P(z_p = 0 \mid \tau_{\setminus p}, \mathbf{Y}_{\setminus p}, \boldsymbol{\Theta}) P(Y_p \mid \mathbf{X}_{\setminus p}, \boldsymbol{\Theta}, z_p = 0) \\ &\propto \alpha P(Y_p \mid \mathbf{X}_{\setminus p}, \boldsymbol{\Theta}, z_p = 0). \end{aligned}$$

If $z_j = 0$ then X_j and τ_j are conditionally independent of $(\mathbf{Y}_{\setminus j}, \tau_{\setminus j})$ and each other given $\mathbf{X}_{\setminus j}$. Let μ_c and σ_c be the conditional mean and standard deviation of $(X_j \mid \mathbf{X}_{\setminus j})$. Therefore, given $\mathbf{X}_{\setminus j}$, the random variable Y_j/σ_c has the same distribution as

$$\frac{\mu_c/\sigma_c + Z}{\sqrt{\tau_j}}$$

for $Z \sim \mathcal{N}(0, 1)$. We recognize the distribution as a noncentral t -distribution with degrees of freedom ν and noncentrality parameter μ_c/σ_c .

Similarly, for $k > 0$,

$$P(z_p = k \mid \tau_{\setminus p}, \mathbf{Y}, \boldsymbol{\Theta}) \propto P(Y_p \mid \tau_{\setminus p}, \mathbf{Y}_{\setminus p}, \boldsymbol{\Theta}, z_p = k).$$

Now, if $z_p = k$, then $Y_p = X_c/\sqrt{\tau_k}$ where $X_c \sim \mathcal{N}(\mu_c, \sigma_c)$. Therefore, the conditional distribution of Y_p given $\boldsymbol{\Theta}, \mathbf{Y}_{\setminus p}, \tau_{\setminus p}$, and $z_p = k$ is

$$f_Y(y) = f_{X_c}(\sqrt{\tau_k}y)\sqrt{\tau_k}.$$

Combining all the above elements gives the result stated in (4.5).

A.3 Full Conditional for Cluster Values in the Dirichlet t -Model

Define $(k) = \{j : z_j = k\}$ and $\setminus(k) = \{1, \dots, p\} \setminus (k)$. First, note that the pair $(\mathbf{Y}_{(k)}, \eta_k)$ is conditionally independent of $\mathbf{Y}_{\setminus(k)}$ given $\mathbf{X}_{\setminus(k)}$. Hence,

$$f(\eta_k \mid \eta_{\setminus k}, \mathbf{Y}, \boldsymbol{\Theta}, z) = f(\eta_k \mid \mathbf{Y}_{(k)}, \mathbf{X}_{\setminus(k)}, z) \propto f(\eta_k, \mathbf{Y}_{(k)} \mid \mathbf{X}_{\setminus(k)}, z).$$

The last density is equal to

$$f(\eta_k \mid \mathbf{X}_{\setminus(k)}) f(\mathbf{Y}_{(k)} \mid \mathbf{X}_{\setminus(k)}, \eta_k, z) = f_{\Gamma}(\eta_k; \nu/2, \nu/2) \cdot f_{\mathcal{N}}\left(\mathbf{Y}_{(k)}; \frac{\boldsymbol{\mu}_c}{\sqrt{\eta_k}}, \frac{\boldsymbol{\Sigma}_c}{\eta_k}\right) \quad (\text{A.3})$$

where $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are the conditional mean vector and covariance matrix of $\mathbf{X}_{(k)}$ given $\mathbf{X}_{\setminus(k)}$. By the well-known formulas for inverse of a partitioned matrix, $\boldsymbol{\Theta}_{(k)(k)}$ is the inverse of $\boldsymbol{\Sigma}_c$, and $\boldsymbol{\mu}_c$ is equal to $\boldsymbol{\Theta}_{(k)(k)} \boldsymbol{\Theta}_{(k)\setminus(k)} \mathbf{X}_{\setminus(k)}$. The product in (A.3) is thus proportional to

$$\begin{aligned} &\eta_k^{\nu/2-1} \exp\{-\eta_k \nu/2\} |\eta_k \boldsymbol{\Theta}_{(k)(k)}|^{1/2} \exp\left\{-\frac{1}{2} \left(\mathbf{Y}_{(k)} - \frac{\boldsymbol{\mu}_c}{\sqrt{\eta_k}}\right)^T \eta_k \boldsymbol{\Theta}_{(k)(k)} \left(\mathbf{Y}_{(k)} - \frac{\boldsymbol{\mu}_c}{\sqrt{\eta_k}}\right)\right\} \\ &= \eta_k^{(\nu + |(k)|)/2-1} \exp\left\{-\eta_k \left[\nu/2 + \text{tr}(\boldsymbol{\Theta}_{(k)(k)} \mathbf{Y}_{(k)} \mathbf{Y}_{(k)}^T)/2\right] - \sqrt{\eta_k} \text{tr}(\boldsymbol{\Theta}_{(k)\setminus(k)} \mathbf{X}_{\setminus(k)} \mathbf{Y}_{(k)}^T)\right\}, \end{aligned}$$

which is the claim of (4.6). Note that when (k) is a singleton, we get the conditional distribution for the alternative model.